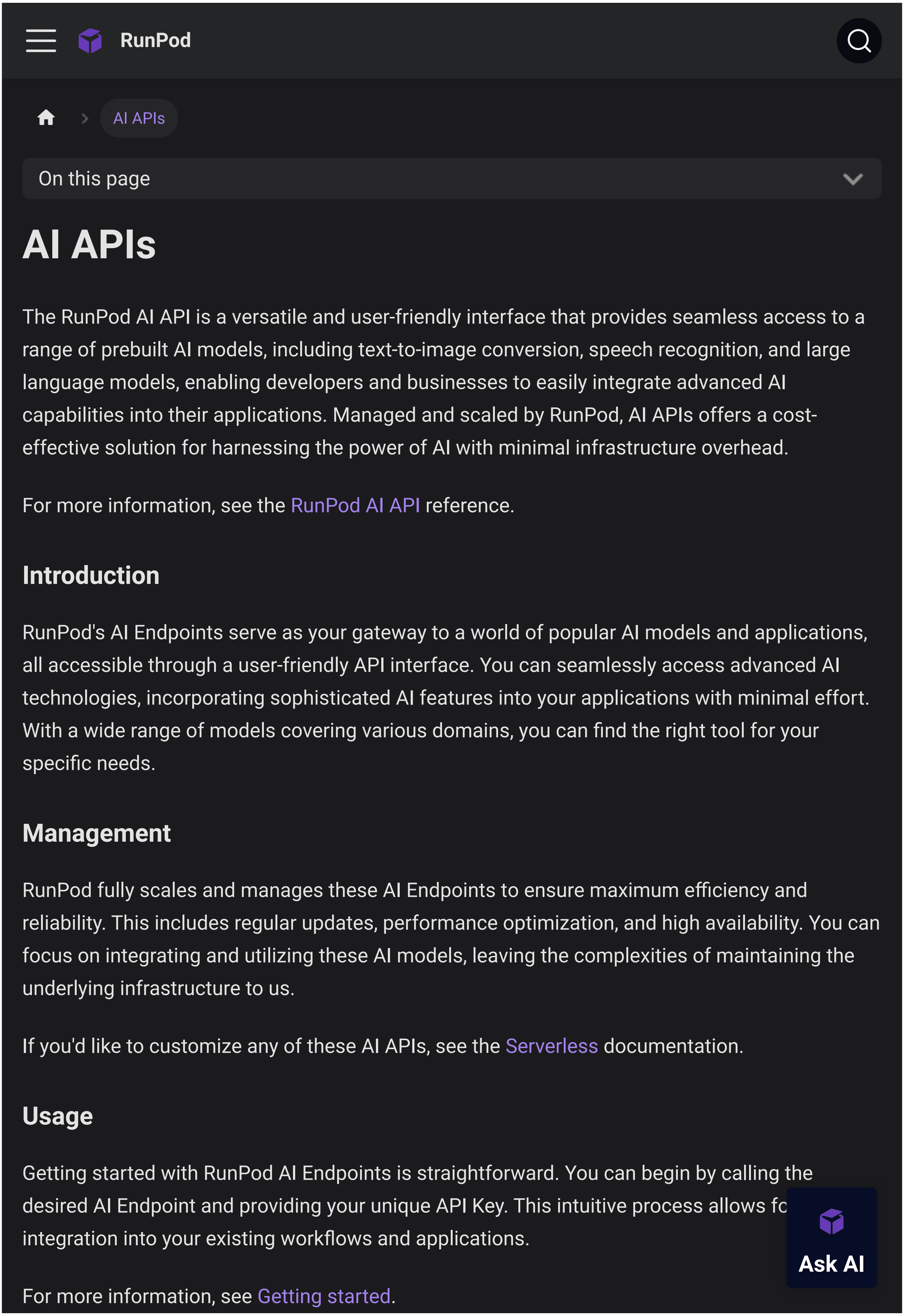


AI APIs

The RunPod AI API is a versatile and user-friendly interface that provides seamless access to a range of prebuilt AI models, including text-to-image conversion, speech recognition, and large language models, enabling developers and businesses to easily integrate advanced AI capabilities into their applications. Managed and scaled by RunPod, AI APIs offers a cost-effective solution for harnessing the power of AI with minimal infrastructure overhead.

For more information, see the RunPod AI API reference.

Introduction

RunPod's AI Endpoints serve as your gateway to a world of popular AI models and applications, all accessible through a user-friendly API interface. You can seamlessly access advanced AI technologies, incorporating sophisticated AI features into your applications with minimal effort. With a wide range of models covering various domains, you can find the right tool for your specific needs.

Management

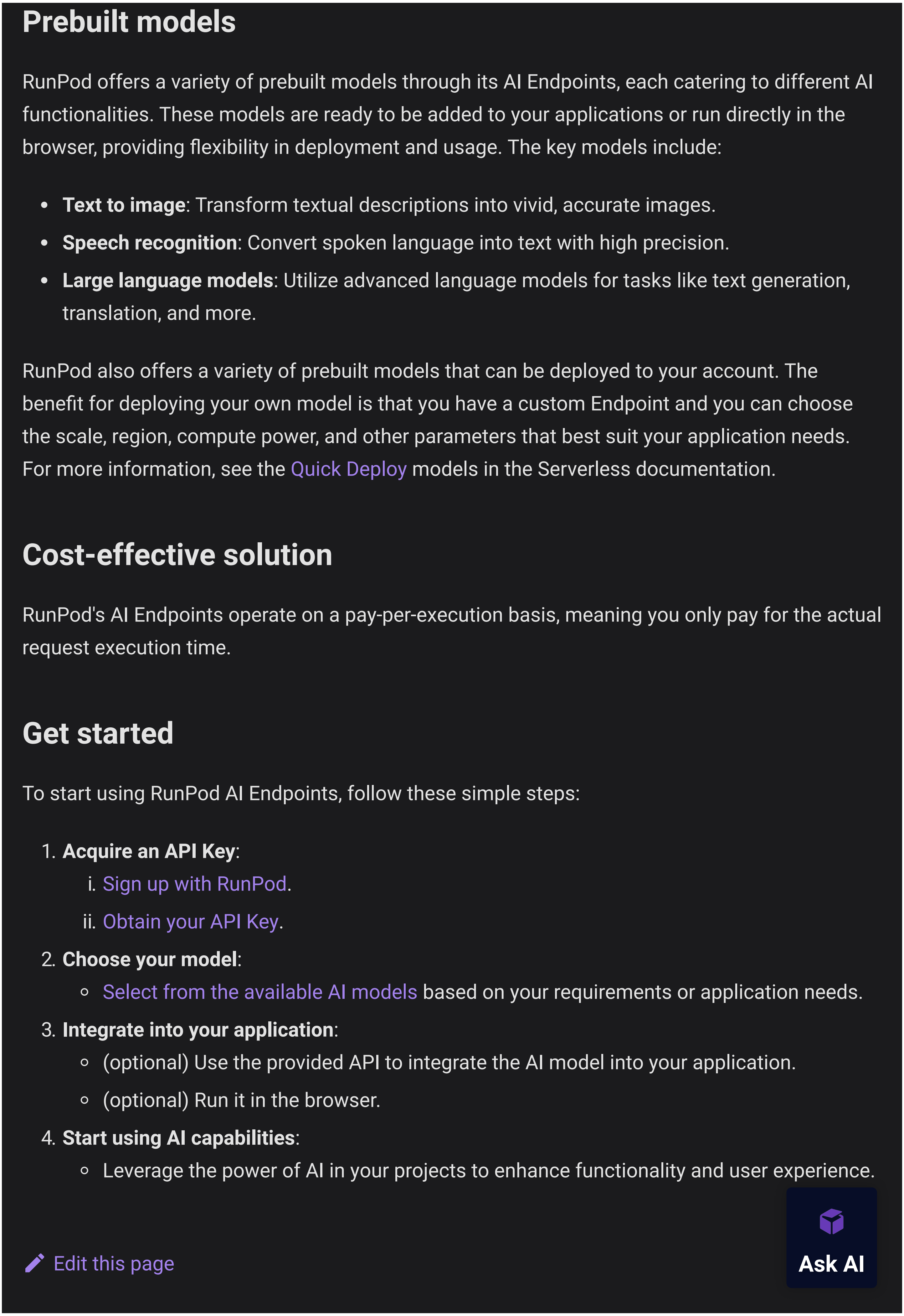
RunPod fully scales and manages these AI Endpoints to ensure maximum efficiency and reliability. This includes regular updates, performance optimization, and high availability. You can focus on integrating and utilizing these AI models, leaving the complexities of maintaining the underlying infrastructure to us.

If you'd like to customize any of these AI APIs, see the Serverless documentation.

Usage

Getting started with RunPod AI Endpoints is straightforward. You can begin by calling the desired AI Endpoint and providing your unique API Key. This intuitive process allows fo integration into your existing workflows and applications.

For more information, see Getting started.



Prebuilt models

RunPod offers a variety of prebuilt models through its AI Endpoints, each catering to different AI functionalities. These models are ready to be added to your applications or run directly in the browser, providing flexibility in deployment and usage. The key models include:

- **Text to image**: Transform textual descriptions into vivid, accurate images.
- **Speech recognition**: Convert spoken language into text with high precision.
- **Large language models**: Utilize advanced language models for tasks like text generation, translation, and more.

RunPod also offers a variety of prebuilt models that can be deployed to your account. The benefit for deploying your own model is that you have a custom Endpoint and you can choose the scale, region, compute power, and other parameters that best suit your application needs. For more information, see the Quick Deploy models in the Serverless documentation.

Cost-effective solution

RunPod's AI Endpoints operate on a pay-per-execution basis, meaning you only pay for the actual request execution time.

Get started

To start using RunPod AI Endpoints, follow these simple steps:

1. **Acquire an API Key**:
   i. Sign up with RunPod.
   ii. Obtain your API Key.
2. **Choose your model**:
   - Select from the available AI models based on your requirements or application needs.
3. **Integrate into your application**:
   - (optional) Use the provided API to integrate the AI model into your application.
   - (optional) Run it in the browser.
4. **Start using AI capabilities**:
   - Leverage the power of AI in your projects to enhance functionality and user experience.

Previous
« Job states

Next
Manage Pods »

Ask AI